

EXPERIMENTS WITH TEMPLATE ADAPTATION IN AN ISOLATED WORD RECOGNITION SYSTEM

F.R. McInnes*, M.A. Jack*, J. Laver*

ABSTRACT

A template-based isolated word recognition system, with adaptation of templates by weighted averaging with recognised input utterances, is described. Experiments with adaptation of speaker-specific and speaker-independent templates are reported. The results show substantial improvements in the recognition accuracies attained. Aspects of interaction between the system and the user are discussed.

INTRODUCTION

The technique of whole-word template matching (ref 1) for isolated and connected word recognition has attained considerable success and has found practical applications for tasks which involve recognition of words from small to medium-sized vocabularies. The systems available mostly employ speaker-specific templates, formed from utterances of the words by the intended user in a training session. Some success has been attained (ref 2) with speaker-independent systems, using several templates for each word of the vocabulary, formed by clustering from utterances by a standard set of speakers.

A shortcoming of the template-matching approach in its basic form is that the templates are derived entirely from the training utterances provided before the start of a recognition session: no use is made of the additional data acquired during the recognition session in the form of recognised input utterances. An adaptation procedure, by which the initial templates are modified progressively to incorporate information from recognised input, can enhance the performance of a template-based speech recognition system by making the templates more truly representative of the user's pronunciations. This is particularly desirable in a system which starts with speaker-independent templates, as the current user's pronunciations may not correspond closely to any of these templates. Adaptation may also help to track gradual changes in the speaker's voice, during an extended recognition session or over a period of days or months.

An isolated word recognition system incorporating a weighted averaging procedure for adaptation of templates is described briefly below (further details may be found in ref 3), and results are reported which show the effects of this adaptation on the accuracy of recognition. Some issues relating to adaptation and the user-system interface are discussed.

DESCRIPTION OF THE SYSTEM

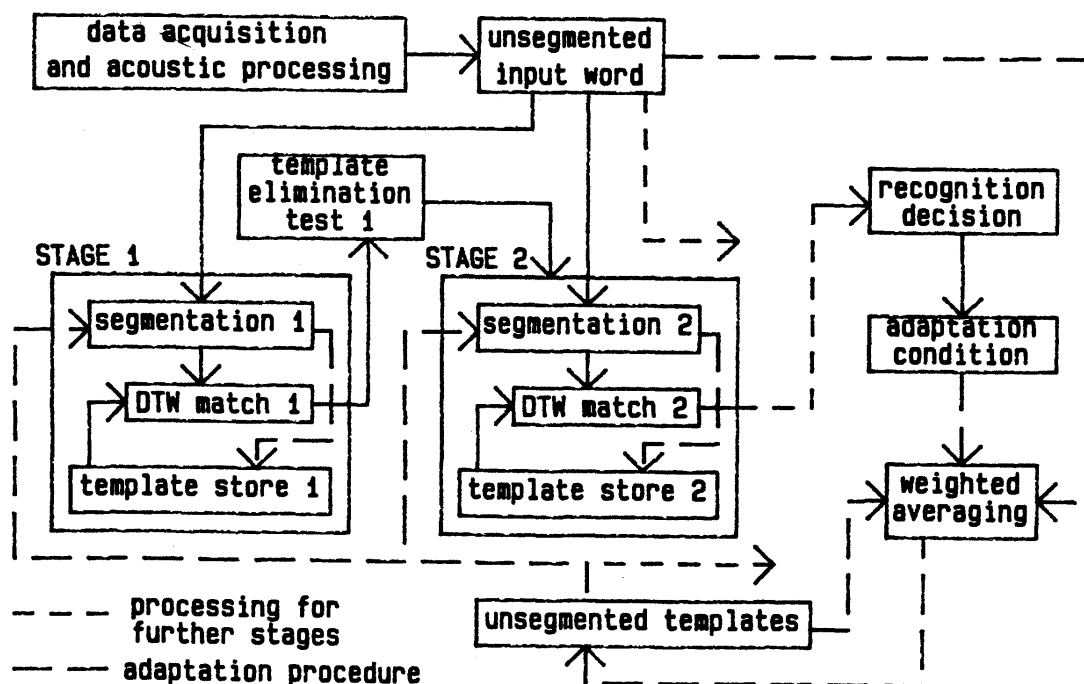
The structure of the isolated word recognition system is illustrated in figure 1. It incorporates a multiple-stage decision procedure, in which successively more detailed comparisons of the input with the templates are carried out until a recognition decision is reached.

For the experiments reported here, three stages were used. Each stage involves division of the input word pattern into a number of equal time segments, and comparison of the resulting normalised pattern with correspondingly segmented forms of the templates by a dynamic programming algorithm (dynamic time warping or DTW) (ref 4). The representation of each word consists of vectors of cepstral coefficients derived from an LPC analysis (ref 5); the segmented form is obtained by averaging these vectors to derive one vector for each segment (or, at the third stage where there are 30 segments per word, interpolating to derive a vector at each segment boundary) (ref 6).

The distances obtained by the DTW comparison at each non-final stage are used to decide which (if any) templates should be matched to the input word at the next stage. If at any comparison stage the ratio of the distances for the best two recognition candidates exceeds a threshold set for that stage, the input is recognised as the word whose template has the smallest distance. Thus the recognition decision may be taken at any of the three stages, depending on whether one word of the vocabulary matches the input much better than any of the others. If the decision can be made after the first stage, the computational cost of the recognition process is very small, as each word is represented at this stage by just two averaged cepstral vectors, and the DTW thus reduces to a very simple linear matching of the input and the template.

*Centre for Speech Technology Research, University of Edinburgh, 80 South Bridge, Edinburgh EH1 1HN

FIGURE 1: STRUCTURE OF RECOGNITION SYSTEM



Once an input word has been recognised, it may be used to adapt the template which has yielded the smallest distance. There are several possible criteria that may be applied to determine whether to perform adaptation.

If there is explicit feedback from the user as to the correctness of the recognition, the condition can be imposed that the recognition must be correct. This case is referred to as *supervised* adaptation. If the recognition is incorrect, the template may be adapted negatively, away from the input word, to make the recurrence of the same misrecognition less likely. It is also possible to test the second-best candidate, where the best is incorrect, and adapt its template also, positively or negatively as appropriate.

Another form of verification of the recognition is possible if the vocabulary includes the special word "CORRECTION". In this case the adaptation to the most recent input word is delayed until the next input is recognised; if this next word is not identified as "CORRECTION", the preceding recognition is assumed to be correct. The indications of correctness or incorrectness obtained by this means can be used to control template adaptation as in the case with explicit verification. The main disadvantage of this option is that wrong adaptations can occur if the word "CORRECTION" is not recognised reliably.

A third form of adaptation condition, which allows *unsupervised* adaptation, does not rely on having any verification of the recognition by the user. The condition imposed in this case is that the ratio of the best two candidates' distances should exceed a threshold value, set to prevent adaptation in cases where the identification of the input is not sufficiently certain.

The adaptation procedure consists of a weighted averaging with DTW alignment (ref 7) applied to the recognised input word and the template to be adapted. The weights on the input and the existing template can be kept constant at successive adaptations, or they can be adjusted so that the ratio of the template weight to the input weight increases linearly with the number of utterances that have gone into forming the template. The former system of weighting is called the *tracking* formulation, because the contribution of each input utterance to the adapted template decays exponentially with subsequent adaptations and so the form of each template depends mainly on the most recent inputs. The latter system is the *optimisation* formulation. Here weights are assigned according to amounts of data, and so an adapted template contains equally weighted contributions from all input utterances used to adapt it.

To improve the stability of the system when the adaptation is unsupervised, a "skewed" adaptation option is provided, for use when there are several templates for each word of the vocabulary (ref 8). The template adapted to any input utterance is not the template with the smallest distance, but the next template in the list for the same word of the vocabulary.

Besides "CORRECTION", two other special words can be included in the vocabulary: "STOP", which, when recognised, causes the termination of the recognition session; and "RETRAIN", which allows

retraining (i.e. formation of a new template to replace the existing one) for any word or words of the vocabulary (which the user selects by keyboard input).

EXPERIMENTS AND RESULTS

Isolated word recognition experiments with template adaptation have been performed using two data bases, with speaker-specific and speaker-independent initial templates respectively.

The data for the speaker-specific template adaptation experiments consisted of words uttered by one male speaker, collected during interactive adaptive recognition sessions with the system described above. The vocabulary, of 50 words, comprised numbers, days of the week and month names. The training and recognition sessions were conducted in a computer terminal room with a moderate but variable level of background noise, using a headset microphone.

Two initial template sets were used, each containing one template for each word in the vocabulary. In template set T1, each template was formed from a single utterance; in set T2, each template was derived by averaging from two utterances of the word. The test data consisted of 10 repetitions of the vocabulary. The numbers of recognition errors occurring on these 10 repetitions are shown in table 1, for cases with and without adaptation. The tracking form of weights was used. (Similar results were obtained with the optimisation form, except that the result using T2 with unsupervised adaptation was improved to 92.8%). In the supervised adaptation case, negative adaptation was employed in cases of misrecognition, but there was no adaptation of the second-best template.

Table 1 Results with adaptation of speaker-specific templates

Adaptation	Errors on repetitions of 50 words										Overall accuracy
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	
(T1)											
None	7	11	8	4	6	2	5	8	5	5	87.8%
Supervised	6	9	7	3	3	2	5	5	3	2	90.0%
Unsupervised	7	10	7	6	6	5	6	9	4	3	87.4%
(T2)											
None	5	5	6	4	8	5	4	7	2	4	90.0%
Supervised	6	4	4	2	1	3	2	4	1	2	94.2%
Unsupervised	6	4	6	2	7	4	4	6	2	2	91.4%

For the speaker-independent recognition experiments, the training data consisted of one repetition of the 10 digits by each of 50 training speakers (37 male and 13 female). Results are given here for two sets of templates, the first (D1) containing six templates for each digit, derived by a criterion based exchange clustering procedure (ref 9), and the second (D2) containing two templates per digit, obtained by separate averaging of the utterances of the male and female training speakers. The test data were three repetitions of the digits spoken by each of 49 speakers (37 male and 12 female) who were not in the training set. The words spoken by each test speaker were recognised, with and without adaptation, using each of the two sets of initial templates. Various forms of adaptation, with the optimisation form of weighting, were tested. A word length normalisation (to 30 vectors per word) by linear time segmentation was applied to all the utterances prior to the clustering and recognition processes.

The results, averaged over the 49 test speakers, are shown in table 2. Average recognition accuracies are given for the first, second and third repetitions of the digits by each speaker, and for the whole set of three repetitions, using each set of initial templates.

Line (1) of table 2 shows the results with no adaptation of the templates. The remaining lines show results with adaptation. The number given after "w" in the left column of the table is the ratio of the weights assigned to the initial template and to each input utterance used in template adaptation. The smaller this ratio is, the faster the adaptation. When it is 0, each adapted template is simply the average of the input utterances used to adapt it.

Lines (2) and (3) show results with supervised adaptation, including negative adaptation (with a small negative weight on the input utterance) for misrecognitions, but not including any adaptation of second-best recognition candidates. Lines (4) and (5) give the corresponding results with second-best candidate adaptation allowed. The improvements over line (1) for the third repetitions show the effect of the adaptation to the preceding two repetitions.

Lines (6) to (8) show results with unsupervised adaptation. The results in line (6) are with adaptation of the best-scoring template; those in lines (7) and (8) are with skewed adaptation. With one exception, these results with unsupervised adaptation show decreases in recognition accuracy.

Table 2 Results with speaker-independent initial templates

Adaptation	Template set D1 Input repetitions				Template set D2 Input repetitions			
	1st	2nd	3rd	all	1st	2nd	3rd	all
None								
(1)	92.4%	94.6%	92.0%	92.99%	91.4%	92.7%	90.6%	91.56%
Supervised								
(2) w1	90.2%	93.9%	94.7%	92.92%	87.8%	91.0%	93.1%	90.61%
(3) w0	90.4%	94.3%	95.3%	93.33%	89.4%	91.4%	94.3%	92.52%
(4) w1 +	89.0%	96.3%	97.8%	94.35%	86.5%	95.7%	97.4%	93.13%
(5) w0 +	89.8%	96.1%	97.6%	94.49%	87.8%	95.7%	97.4%	93.54%
Unsupervised								
(6) w4	91.4%	92.4%	91.4%	91.77%	91.0%	90.0%	88.4%	89.80%
(7) w4 skew	92.0%	93.1%	92.0%	92.38%	90.8%	92.0%	89.0%	90.61%
(8) w2 skew	92.0%	91.6%	90.8%	91.77%	91.0%	90.0%	88.4%	89.80%

The recognition of each speaker's first repetition of the digits is consistently poorer with adaptation than without. This occurs because, during recognition of the first repetition of the vocabulary, the template set is a mixture of unadapted and adapted templates; an adapted template for an incorrect candidate recognition may be closer to the input word than the unadapted correct-candidate template, because adapted templates correspond better to the speaker's voice.

DISCUSSION AND CONCLUSIONS

It is evident from the results obtained that supervised adaptation of templates during recognition sessions can significantly improve isolated word recognition accuracy, whether the initial templates are speaker-specific or speaker-independent. Moreover, the improvement is attained more rapidly, at least with speaker-independent initial templates, if adaptation can be applied not only to the best-matching template but also, where the best candidate is incorrect, to the second-best.

The results with unsupervised adaptation are less consistent. It yielded a net improvement in results with speaker-specific templates, but a deterioration with speaker-independent templates - though in the case of skewed adaptation for multiple templates (D1) the results are not conclusive, and experiments with more extended input sequences will be required to determine whether this adaptation is beneficial.

In the system described here, template adaptation improves not only the accuracy but also the speed of recognition, because, when the templates are well tuned to the speaker, fewer comparisons are required at the later stages of the decision procedure. However, the adaptation itself takes some computing time - often more than the actual recognition. This computation could be reduced by using a linear averaging operation instead of the DTW method.

The design of the interaction between the recognition system and the user is important. By including a convenient means for the user to correct wrong recognitions, and delaying the adaptation to each input until an opportunity for such correction has been given, supervised adaptation can be implemented without the need for an explicit yes/no response by the user to each recognition. The facility for retraining templates as required is a desirable feature, particularly if there is any risk of instability arising from adaptation to inputs which are misrecognised or affected by noise.

ACKNOWLEDGEMENT

The work reported here was made possible by an SERC research studentship.

REFERENCES

1. L R Rabiner & S E Levinson, *IEEE Trans Commun* **COM-29**, 621 (1981)
2. L R Rabiner, S E Levinson, A E Rosenberg & J G Wilpon, *IEEE Trans Acoust, Speech, & Signal Process* **ASSP-27**, 336 (1979)
3. F R McInnes, M A Jack & J Laver, *Proc Inst of Acoust* **8**, 7, 283 (1986)
4. F Itakura, *IEEE Trans Acoust, Speech, & Signal Process* **ASSP-23**, 67 (1975)
5. A H Gray & J D Markel, *IEEE Trans Acoust, Speech, & Signal Process* **ASSP-24**, 380 (1976)
6. F R McInnes, M A Jack & J Laver, *IEE Conf Pub* 258 (SIOTA 86), 21 (1986)
7. R Zelinski & F Class, *Proc IEEE ICASSP* **83**, 1053 (1983)
8. T R G Green, S J Payne, D L Morrison & A Shaw, *Behaviour & Inf Tech* **2**, 1, 23 (1983)
9. A Mokeddem, H Hugli & F Pellandini, *Proc IEEE-IECEJ-ASJ ICASSP* **86**, 2691 (1986)